ArguBias: Quantifying the Impact of Semantic-Positional Misalignment on Argument Similarity

Anonymous ACL submission

Abstract

Despite NLP advances, computational approaches for judging argument similarity face a fundamental challenge: semantic-positional dissonance. Embedding models must distinguish between arguments sharing similar linguistic characteristics yet advancing opposing positions, and conversely recognizing when diverse linguistic expression across different cultural, societal, and philosophical contexts convey identical positions. This distinction between content, rhetoric, and position is a complex issue that requires insight from both cognitive science and computational social science. To address this challenge, we introduce ArguBias, a novel framework that systematically identifies, evaluates, and improves similarity judgments for arguments containing cognitive bias structures. First, we introduce the ArguBias Corpus, containing 8,000 annotated argument pairs facilitating the taxonomy of previously unexamined cognitive bias structures in argumentation. This allows us to benchmark 10 state-of-the-art embedding models on their cognitive bias vulnerability. Finally, we demonstrate how minimal fine-tuning on the ArguBias corpus reduces vulnerability of embedding models to cognitive bias structures by up to 11.6pp. Simultaneous gains of 7.1% and 5.4% on argument similarity benchmarks BWS and AFS indicate generalizability and improvement of fundamental semantic understanding beyond domain-specific applications.

1 Introduction

Why do embedding models incorrectly assess argument similarity? Embedding models distill natural language into numerical representations, enabling computational linguistic analysis. These models learn from vast human-labeled datasets, which may unintentionally transmit our judgment errors into their foundational understanding (Bender and Friedman, 2018; Sap et al., 2022). Humans frequently rely on heuristics that lead to systematic errors called cognitive biases (Berthet, 2022), such as the availability heuristic, anchoring, and framing. In argumentation, these **cognitive bias structures** manifest as the presence or absence of patterns that systematically affect similarity perception, often conflating the semantic and positional dimensions of argument alignment.[1] For example, arguments with similar emotionally-charged language might be perceived as sharing the same perspective despite advocating opposite positions.

Transformer-based sentence embeddings (like S-BERT) which have advanced NLP performance significantly, are fundamentally optimized to cluster texts with similar meaning in vector space. This semantic similarity might not be sufficiently optimized or appropriate for stance-dependent tasks (Hanley and Durumeric, 2024; Joshi et al., 2020; Barak et al., 2019) such as opinion mining (Ghafouri et al., 2024). Embedding models risk overemphasizing semantic overlap, deviating from the expected judgment of positional opposition. Effective argument similarity detection must balance these competing factors, recognizing when divergent rhetoric masks equivalent positions and when shared language conceals fundamental disagreement.

The discrepancy between semantic and positional similarity has serious practical implications beyond theoretical concerns. When semantic similarity models misidentify positional alignment, they can undermine critical applications like misinformation detection and political ideology similarity. (Pennycook and Rand, 2021) finds that heuristics, such as confirmation bias, contribute to the spread of misinformation. Studies relying on embeddings to study (dis)similarity between political party ideologies are potentially prone to flawed alignment

[1] This is a shift from the traditional paradigm of cognitive biases from an internal cognitive process to an observable pattern between two arguments that causes distortions in decision-making. See Section 1.1 & 3.2 for more.

measurements (Ceron et al., 2022). The fundamental issue is the systematic overemphasis on semantic patterns, conflating argumentative position with semantic similarity.

Contribution. In this work, we develop a framework at the intersection of cognitive and computational social science, advancing the understanding of positional similarity of arguments for a variety of downstream tasks. The contributions of our work are the following:

a. The development of a novel taxonomy of cognitive bias structures in argumentation. We identify 10 patterns in argumentative texts that can lead to a system deviating from the normative judgment decision of similarity.

b. The release of a novel argumentation dataset containing aforementioned cognitive biases. We collect and annotate the ArguBias Corpus, a dataset of nearly 8,000 matched arguments from US Congressional committee hearings and Reddit discussions, to be released upon acceptance (CC-BY 4.0) to advance research on cognitive biases in computational argumentation.

c. The comprehensive benchmarking of modern text embedding models on their cognitive bias vulnerability. Utilizing the ArguBias Corpus, we quantify the performance and bias-specific vulnerabilities of proprietary and open-source text embedding models when judging similarity of arguments with cognitive bias structures.

d. Mitigation techniques for improving embedding vulnerability. Finally, we demonstrate how minimal fine-tuning of embedding models on the ArguBias Corpus (1) reduces the embedding models' vulnerability to cognitive biases and (2) improves their performance on argument similarity benchmarks.

Key Results. In this work, we find that (1) text embedding models exhibit a critical vulnerability to lexical overlap bias, suffering severe similarity overestimation 2-4 times worse than resistant biases, revealing that shared vocabulary misleads argument similarity judgments more severely than structural argumentative reasoning patterns.

We show that (2) embedding models experience a 14.3% drop in positional similarity detection performance when evaluating argument pairs made exclusively in congressional hearings versus those made on Reddit.

Finally, we demonstrate how (3) fine-tuning on the ArguBias corpus, though containing only four topics, improves the baseline performance on unseen AFS and BWS [2] argument similarity (+5.4% and +7.1%). This suggests embedding models can learn domain-agnostic cognitive bias structure mitigation, enabling more robust and transparent computational argumentation tools.

Bias Type	Platform Pairs		
	C→R	R→R	C→C
Framing bias	650	1,077	1,637
Lexical overlap	896	1,183	607
Linguistic distance	235	171	69
Affect heuristic	137	315	155
Confirmation bias	84	198	71
Availability heur.	55	109	153
Anchoring bias	22	72	31
None	3	13	18
Implicit association	6	2	5
Halo effect	2	1	6
False equivalence	1	7	1
Total	*2,091*	*3,148*	*2,753*

Table 1: Distribution of cognitive bias types across platform argument pairs (Congress→Reddit, Reddit→Reddit, Congress→Congress).

1.1 Paper Scope

Critically, we distinguish our use of 'cognitive bias' from its colloquial meaning, focusing solely on systematic deviations from rational judgment standards that can help inform model improvements (Tversky and Kahneman, 1974). Our paper employs cognitive bias structures not as a comprehensive explanation of misbehavior, but as a framework within post-hoc interpretability research to provide meaningful insights into black-box decision-making (Oh, 2024).

2 Motivation & Related Work

Our work is motivated by the difficulty and disconnect in the literature on quantifying argument similarity computationally. Embedding approaches have revolutionized Natural Language Processing, but it faces issues in the application of nuanced argument modeling. Specifically, similarity measures have so far focused primarily on mathematical representations and vector comparisons, but these methods do not always adequately capture human cognitive processes (Barak et al., 2019).

The need for accurate argument modeling requires a comprehensive approach grounded in a cognitive science framework, which is the goal of

[2] These benchmarks are explained in Section 4.3 & 4.3

Model	Cognitive Bias Vulnerability (Brier Score)										
	AV	FR	CO	AF	IA	LO	AN	LD	FE	HE	NO
BGE-v1.5	0.090	0.086	0.268	0.100	0.182	0.526	0.146	0.074	0.490	0.190	0.057
MiniLM-v2	0.228	0.230	0.247	0.237	0.222	0.299	0.205	0.218	0.214	0.227	0.105
MPNet-v2	0.192	0.206	0.243	0.217	0.231	0.307	0.198	0.205	0.282	0.220	0.090
MultiMPNet	0.178	0.178	0.243	0.183	0.220	0.341	0.189	0.175	0.305	0.178	0.075
Voyage-3	0.100	0.108	0.255	0.110	0.187	0.451	0.149	0.099	0.428	0.181	0.051
Nomic-v1.5	0.048	0.045	0.301	0.061	0.195	0.643	0.141	0.036	0.595	0.187	0.016
Jina-v3	0.207	0.213	0.233	0.207	0.236	0.285	0.193	0.206	0.248	0.183	0.103
Mistral	0.040	0.035	0.319	0.052	0.190	0.683	0.139	0.027	0.670	0.184	0.015
Embed-v3	0.171	0.174	0.230	0.195	0.212	0.310	0.173	0.178	0.284	0.189	0.103
OpenAI	0.238	0.251	0.231	0.262	0.256	0.226	0.212	0.250	0.200	0.248	0.144

Table 2: Vulnerability of embedding models across cognitive bias types (AV: Availability, FR: Framing, CO: Confirmation, AF: Affect, IA: Implicit Association, LO: Lexical Overlap, AN: Anchoring, LD: Linguistic Distance, FE: False Equivalence, HE: Halo Effect, NO: None). Brier scores measure calibration quality, with lower values indicating better performance. Colors range from light gray (best) to red (worst).

our work. The first step is to understand and categorize the errors in current computational argumentation approaches. Recently, there has been work on measuring cognitive biases in LLM evaluation (Koo et al., 2024; Malberg et al., 2024). However, these works do not address the role of cognitive bias structures in computational argumentation nor their effect on establishing argument similarity.

2.1 Cognitive Biases in Computational Reasoning

Research in cognitive science has developed a comprehensive taxonomy of systematic cognitive biases in human reasoning that affect rational thought (Stanovich et al., 2008; Baron, 2023). These biases have been thoroughly analyzed to understand their impact on suboptimal decision-making, strategic planning, and information collection (Barnes Jr, 1984; Joyce and Biddle, 1981). Studies have looked at the adversarial impact of specific cognitive biases (Zhdanko, 2019), citing emotional resonance (affect heuristic) as a significant source of manipulation of young students online.

Limited research directly addresses the problem as framed here. Recent advances in our understanding of the behavior of complex LLMs [3] (Malberg et al., 2024) have revealed the significant vulnerability of these models to bias effects. However, we have not found studies that explore these issues specifically within the computational argumentation context.

The investigation of cognitive biases is traditionally done in computational text analysis under a systematic behavioral lens (Venkit and Wilson, 2021), where cognitive biases are seen as emergent properties of statistical learning. Our work applies classical decision judgment paradigms (Tversky and Kahneman, 1974) to instead view cognitive biases as systematic deviations from normative decision standards.

2.2 Argument Similarity

Several efforts show the difficulty of assessing argument and stance similarity.

A recent study (Ghafouri et al., 2024) showcases the utility of fine-tuning sentence transformers, examining the role of similar hashtags and terminology as illusions of false similarity. Another study (de Sousa and Becker, 2023) identifies the risk of over-reliance on semantic similarity of hashtags for the retrieval of tweets for stance, resulting in false positives. The task of same-side stance classification emerged as an alternative to the substantial requirement for domain knowledge in stance classification. A study (Körner et al., 2021) found that a model that only considers linguistic similarity fails to handle oppositional cases. This over-reliance on domain-dependent semantic similarity provides significant challenges for topic-agnostic argument similarity analysis.

Similar challenges appear in computational argumentation for deliberation analysis. A recent study (Plenz et al., 2024) highlights the importance of finer-grained evaluation of argument similarity, as similar issues share similar concepts and framing. This is especially important for accurate cross-domain deliberation comparison studies (Irani et al., 2025a). There is also an interest in the explainability of stance predictions: a recent study (Saha et al., 2024) builds a stance tree utilizing

[3] LLMs use embeddings as the building blocks for complex language generation and understanding tasks.

rhetorical parsing to construct an evidence tree.

Overall, these works emphasize the need and importance for explainable and robust computational argumentative positional assessments.

3 Definitions

In this section, we outline and define: (a) two aspects of similarity, and (b) cognitive bias structures that play a critical role in our framework.

3.1 Semantic and Positional Similarity

Semantic similarity refers to the degree of likeness in meaning or semantic content between two pairs of arguments. **Positional Similarity** refers to the degree of alignment between the viewpoints, stances, or perspectives of two arguments. Position, in argumentation theory, encompasses the colloquial meaning of stance, with the added requirement of an argumentative structure (presence of a claim supported by at least one premise).

3.2 Cognitive Bias Structures

Our framework draws on cognitive biases identified in established literature, specifically (Nickerson, 1998; Tversky and Kahneman, 1973, 1974; Greenwald et al., 1998; Zajonc, 1968; Tversky, 2003; Slovic et al., 2007; Tversky and Kahneman, 1981; Kaster et al., 2021). We selected the following biases through a systematic analysis of computational argumentation challenges, prioritizing those most directly affecting similarity judgments while also representing a diverse set of cognitive mechanisms.

3.3 Surface vs. Structural Biases

Surface-level bias patterns, including availability, lexical overlap, linguistic distance, and implicit association operate directly on the precise wording, ordering and formatting of two arguments. In contrast, structural biases such as framing, confirmation, affect, false equivalence and anchoring impact the perception of how each claim is *linked* to its supporting premises. Therefore, when structural bias patterns are present, deciding whether two arguments adopt the same or differing position often requires explicit understanding or analysis of this relational structure (Jullien, 2016). Surface-level bias patterns, on the other hand, impact similarity decisions without affecting claim-premise relationships (McCoy et al., 2019).

1. Confirmation Bias refers to the distortion in how systems judge argument alignment by either overemphasizing shared ideological markers (political terminology, value-laden terms, identity based vocabulary) despite substantive differences in positions. Or underestimating similarity between arguments containing similar positions when they employ different or opposing ideological markers.

2. Availability Heuristic refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments with easily recalled features such as vivid language or concrete examples, or underestimating similarity between arguments when the degree of vividness differs despite similar positions.

3. Anchoring Bias refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that share initial elements (opening claims, premises, examples), while overlooking significant positional divergence thereafter. Or underestimating similarity between arguments with different initial elements despite their convergence on substantively similar positions.

4. False Equivalence Bias refers to the distortion in how systems judge argument alignment by either overemphasizing shared argumentative presentation similarities (shared reasoning pattern, e.g., If $A \rightarrow B$ therefore B). Or underestimating similarity between arguments employing different reasoning patterns but having similar positions.

5. Implicit Association refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that share group-associated language (cultural references, gendered terminology, political signals) despite advancing different positions. Or underestimating similarity between arguments that use different group associated language despite advocating substantively similar positions.

6. Linguistic Distance Bias refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that share linguistic characteristics (vocabulary complexity, formality level), despite addressing different positions. Or underestimating similarity between arguments expressing similar positions when they employ different linguistic styles.

7. Affect Heuristic refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that contain comparable emotional characteristics (shared positive tone, outrage responses) despite addressing different positions. Or underestimating similarity between arguments containing similar positions

when they evoke contrasting emotional tones.

8. Framing Bias refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that share contextual framing approaches (similar moral language, legal framing) despite advocating for different positions. Or underestimating similarity between arguments expressing similar positions when they employ different framing techniques.

9. Lexical overlap bias refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments that share terminology or vocabulary (matching technical terms or jargon) despite fundamental differences in positions. Or underestimating similarities between arguments expressing similar positions when they employ different terms or lexicons.

10. Halo Effect refers to the distortion in how systems judge argument alignment by either overemphasizing similarities in arguments with comparable authority or credibility characteristics (prestigious affiliations, credentials, references) despite containing different positions. Or underestimating similarity between arguments with equivalent positions when they come from sources with differing levels of perceived authority or credibility.

4 Data

Evaluating cognitive biases in computational argumentation remains unexplored. To address this, we create a novel dataset, the ArguBias Corpus, containing nearly 8,000 pairs of short-form argumentative texts from US Congressional committee hearings and Reddit discussions, annotated with binary positional similarity labels and cognitive bias structures if present. We focus on four controversial policy topics: *Abortion*, *Gun Control*, *Nuclear Energy* and *GMOs*. Dataset characteristics are summarized in Table 1, with data collection and compilation detailed in the following section.

4.1 The ArguBias Corpus

We extracted approximately 100,000 publicly available verbal statements from U.S. House committee hearings (2005-2022) across four policy domains: Abortion (37 hearings), Gun Control (32), GMOs (13), and Nuclear Energy (316). Publicly accessible Reddit discussions were selected based on community activity, moderator descriptions, and manual relevance verification. Both congressional and social media sources underwent iterative keyword expansion to capture diverse policy terminology and separate PII redaction.

Model	ROC-AUC	PR-AUC	Dim.
Embed-v3	**0.6337**	**0.7516**	1024
OpenAI	0.6260	0.7325	3072
Mistral	0.6053	0.7102	1024
Jina-v3	0.5877	0.7028	1024
MPNet-v2	0.5803	0.6847	768
Voyage-3	0.5798	0.6906	1024
MultiMPNet	0.5764	0.6916	768
Nomic-v1.5	0.5743	0.6894	768
MiniLM-v2	0.5612	0.6699	384
BGE-v1.5	0.5265	0.6562	1024

Table 3: Comparison of Embedding Models' Cognitive Bias Performance by ROC-AUC, PR-AUC, and Dimensionality.

We apply BERTopic (Grootendorst, 2022) and RAKE (Rose et al., 2010) keyword expansion on the initial selection of hearings and Reddit posts, and collect the outputted keywords for each policy issue. Based on the expanded set of keywords, we select additional Reddit posts and hearings and apply BERTopic and RAKE once more to finalize a set of keywords. For hearings, we take additional screening steps and select the hearings in which these keywords appear more than three times, and then we took the first 3,000 words of these hearings and asked a LLaMa 3.1 8B model with a low temperature to identify whether each hearing is primarily about the given policy issue. We select Reddit posts that contain at least one of the keywords.

Argument & Topic Extraction. To extract arguments from our corpus of committee hearing statements and Reddit posts, we use WIBA, an open-source argument detection and extraction tool capable of identifying arguments within short or long-form text (Irani et al., 2025b). This tool has been shown to have high performance on argument detection benchmarks, with $F1$ scores of above 80%. Furthermore, WIBA was shown to be particularly effective at detecting arguments written with different levels of formality which directly aligns with our needs. WIBA is also capable of extracting the topic being argued, which is how we filter and identify appropriate arguments. This leaves us with 85,325 extracted Reddit arguments and 47,252 committee hearings arguments across our 4 issues.

Argument Pair Selection. To identify and analyze pairs of arguments from different sources, we focus on finding potential missed and false similarity matches. We assume that arguments with similar positions often anchor around the same named

Model	C-C Mean	C-C Std	R-R Mean	R-R Std	C-R Mean	C-R Std	d
MiniLM-v2	0.56	0.17	0.54	0.17	**0.51**	**0.20**	0.22
MPNet-v2	0.59	0.15	0.55	0.16	**0.52**	**0.17**	0.33
BGE-v1.5	0.72	0.08	0.74	0.06	**0.71**	**0.07**	0.27
Embed-v3	0.60	0.08	0.58	0.08	**0.55**	**0.09**	0.47
Jina-v3	0.56	0.12	0.56	0.12	**0.51**	**0.12**	0.43
Mistral	0.85	0.04	0.84	0.03	**0.83**	**0.04**	0.42
OpenAI	0.53	0.11	0.49	0.10	**0.46**	**0.11**	0.41
MultiMPNet	0.61	0.14	0.59	0.13	**0.56**	**0.15**	0.27
Voyage-3	0.68	0.08	0.71	0.08	**0.65**	**0.09**	0.54
Nomic-v1.5	0.82	0.06	0.81	0.06	**0.80**	**0.07**	0.32
Average	0.65	0.10	0.64	0.10	**0.61**	**0.11**	0.37

Table 4: Cosine similarity for within-platform (C-C, R-R) versus cross-platform (C-R) argument pairs. Cross-platform similarities are always lower, indicating source-specific bias (all $p < 10^{-15}$). Cohen's d shows effect sizes for same-source vs. cross-source comparisons.

entities, even if their rhetoric or linguistic structures diverge. While opposing arguments may reference the same entities but emphasize different aspects of these common reference points. Furthermore, by selecting topics that are rooted in modern policy discourse we can assume that specific legislation, organizations and geographic entities will serve as common reference points for positional perspectives. Examples of these entities include Roe v. Wade (Abortion legislation), Planned Parenthood (Abortion organization), and Schools (Gun Control location). Additionally, we use Jaccard similarity between entity sets to capture potential missed matches when exact entity overlap is absent.

We define two categories of argument pairs: potential missed similarities and potential false similarities. Potential *missed* similarity pairs must have a high entity overlap, ≥ 3 matched entities or a Jaccard similarity ≥ 0.3, as well as a semantic similarity ≤ 0.5. These represent cases where embeddings may fail to recognize semantic similarity. Potential *false* similarity are calculated by finding argument pairs where there is low entity overlap, ≤ 0.1 Jaccard similarity and high semantic similarity ≥ 0.7. This represents cases where embedding models may incorrectly assume similarity.

4.2 Annotation Process

We validated our ArguBias dataset through a multi-stage process, beginning with prompt-engineered ChatGPT o3 annotations, followed by manual human verification. We sampled 200 argument pairs using stratified sampling that deliberately focused on challenging boundary cases: 80 false positives, 80 false negatives, and 40 true positives/negatives. This approach ensured representation across all cognitive bias types, while concentrating annotation effort on the most informative and difficult cases. This strategy naturally increases annotation complexity but provides stronger validation of our schema's reliability. Four trained undergraduate public policy students annotated these samples, for course credit, using the exact framework for positional similarity and cognitive biases in Sections 3.1 & 3.2. We achieved moderate inter-annotator agreement (Cohen's $\kappa = 0.471$) on positional similarity judgments and 95% agreement on bias structure identification. This level of agreement is consistent with other complex discourse annotation tasks requiring semantic inference and subjective judgments (Brambilla et al., 2022; Lawrence and Reed, 2020; Hoek et al., 2021). We continue to evaluate additional samples to further improve reliability.

4.3 Benchmark Datasets

We use the Argument Facet Similarity (AFS) corpus to evaluate argument similarity improvement as a result from fine-tuning on the ArguBias dataset. The AFS corpus contains 6,000 argument pairs from 3 topics, gun control, gay marriage, and the death penalty (Misra et al., 2017). Each pair is annotated with a similarity scale from 0 to 5, with 5 indicating equivalence and 0 indicating different topics. The AFS corpus is used in the evaluation and is never used in fine-tuning.

The BWS Argument Similarity Corpus contains 3,400 argument pairs covering 8 controversial topics, cloning, abortion, minimum wage, marijuana legalization, nuclear energy, death penalty, gun control, and school uniforms, with 425 pairs per topic (Thakur et al., 2020). This dataset is also only used an indicator for argument similarity task improvement as a result of fine-tuning.

5 Experiment Setup

This section outlines our experiment setup for (1) benchmarking modern text embedding models against ArguBias and (2) fine-tuning embedding models to improve cognitive bias vulnerability.

5.1 Embedding Models Selection

We use the most recent leaderboard for MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2022) provided by HuggingFace[4] and

[4]https://huggingface.co/spaces/mteb/leaderboard

filter for models that have performed best for STS (Semantic Textual Similarity) and MTEB Score.

Open-source models. Our evaluation framework incorporates diverse open-access embedding architectures: *MPNet-base-v2*, *MiniLM-L6-v2*, *BGE-large-en-v1.5*, and *Multilingual-MPNet*. We analyze these models in zero-shot configurations without any task-specific fine-tuning to establish critical performance benchmarks. This approach reveals inherent representational capacities across varying model scales and architectural designs when processing argument similarity.

Proprietary models. Our evaluation includes leading proprietary embedding models: *Embed-english-v3.0* (Cohere), *OpenAI*, *Mistral-embed*, *Jina-embeddings-v3*, and *Voyage-3-large*. These state-of-the-art models demonstrate exceptional performance in general semantic similarity tasks. We leverage their advanced representational capabilities to examine how commercial embedding technologies capture the nuanced semantic-positional relationships in argumentative contexts.

5.2 Evaluation Metrics

Discrimination Metrics. We measure models' ability to rank similar argument pairs higher than dissimilar pairs despite cognitive bias interference. **ROC-AUC** quantifies the probability that a randomly selected similar pair ranks higher than a dissimilar pair, providing separability assessment regardless of class distribution. **PR-AUC** evaluates precision-recall performance, particularly sensitive to model behavior in imbalanced scenarios. Values approaching 1 indicate preserved ranking ability under bias conditions. For ROC-AUC, values near 0.5 suggest random performance, indicating that cognitive bias structures disrupt discrimination.

Vulnerability Assessment. We use the **Brier Score (BS)** as the primary performance measure for bias-specific evaluation. Since argument similarity constitutes binary classification, Brier score equals the mean squared error between predicted similarity scores and ground truth labels. Lower Brier scores indicate reduced vulnerability to cognitive bias structures, with perfect performance at 0 and worst possible performance at 1 (Brier, 1950).

5.3 Fine-tuning

We fine-tune three open-access embedding models (MPNet-v2, Nomic-v1.5, and Jina-v3) for 2 epochs, on ArguBias with a 70/15/15 train/dev/test split. All models use Online Contrastive Loss with Matryoshka Loss function (Kusupati et al., 2024) and a learning rate of $5e{-}5$. These parameters were determined through hyperparameter search. Matryoshka Representation Learning has been shown to be particularly effective in the training of OpenAI's embedding model *text-embeddings-3-large* (Tamber et al., 2024), which is also a high performer against ArguBias. This approach enables models to learn from the most challenging positive and negative argument pairs, while optimizing representations across multiple embedding dimensions.

Benchmark	**Jina-v3**		**MPNet-V2**		**Nomic-v1.5**	
	Base	**Fine**	**Base**	**Fine**	**Base**	**Fine**
$\text{ArguBias}_{\text{Test}}$	58.3	65.1	58.9	**70.5**	57.9	66.2
AFS†	55.7	58.3	53.9	58.8	58.0	**59.4**
BWS†	53.3	56.3	53.8	**59.3**	56.1	59.2
Bias-Specific (BS)						
Framing[504]	11.8	12.0	13.5	11.5	11.3	**10.3**
Lexical Overlap[404]	42.8	41.6	**39.1**	42.2	44.1	45.9
Affect Heuristic[92]	12.8	13.0	14.7	12.4	12.4	**11.5**
Confirmation[53]	25.5	25.1	**24.7**	25.0	25.9	26.2
Linguistic Distance[71]	11.5	11.9	13.0	11.3	10.7	**9.8**
Availability[48]	11.7	11.8	13.0	11.1	11.4	**10.4**
Anchoring[19]	16.0	16.0	16.5	**15.2**	15.7	15.3

Table 5: ArguBias fine-tuning results. ArguBias Test: ROC-AUC scores (%). †AFS/BWS use cosine Spearman correlation. Bias-specific: Brier scores (lower is better). Superscripts show sample sizes. Bold indicates the best performance.

6 Findings

6.1 Quantifying Embedding Vulnerability

Finding 1: Cognitive biases degrade embedding models' argument similarity judgments through distinct mechanisms.

Analyzing ROC-AUC and Brier Scores across sampled bias types ($n \geq 100$) reveals three vulnerability patterns. Models demonstrate strong resistance to linguistic properties (linguistic distance bias, availability heuristic) and argumentative framing (ROC-AUC > 0.7 , BS < 0.25). This suggests effective judgment of argument positional similarity despite differences in linguistic style, vividness, or presentation approach.

Lexical overlap creates a critical ranking-calibration disconnect. Models preserve ranking ability while suffering similarity calibration 3-4 times worse compared to resistant biases (linguistic distance, availability, framing). This overconfidence when arguments share vocabulary aligns

with existing NLP research (Rajaee et al., 2022).

Other structural biases (affect heuristic, confirmation bias, anchoring bias) demonstrate 15-35% performance degradation across both metrics, with anchoring bias showing the most severe impairment. Rare structural biases ($n < 100$) exhibit severe performance degradation, but require further investigation due to insufficient statistical power.

These findings reveal that state-of-the-art embedding models are fundamentally vulnerable to being systematically misled by surface-level lexical overlap. This vulnerability is more severe than complex structural biases affecting claim-premise relationships (see Table 2).

6.2 Cross-Platform Performance Degradation

Finding 2: Embedding models demonstrate performance degradation across platform boundaries, with positional similarity detection varying by argumentative context.

Models exhibit a 5.3% ROC-AUC degradation when transitioning from Congress-to-Congress to Congress-to-Reddit comparisons. Performance drops further to 14.3% below the Congress-Congress baseline for Reddit-to-Reddit pairs. Proprietary models (Embed-v3, OpenAI, Voyage-3) show the highest sensitivity to platform shifts.

Similarity performance varies by platform combination. Cross-platform evaluation improves ROC-AUC scores by 9.5-26.6% when argument pairs are affected by lexical overlap, anchoring, and framing biases, while decreasing performance by 4.5-12.7% for pairs containing confirmation bias, availability heuristic and affect heuristic structures. Congress-to-Congress pairs achieve 50.4% higher ROC-AUC for affect heuristic affected pairs and 27.4% higher for confirmation bias-affected pairs than Reddit-to-Reddit comparisons.

Source-specific representational bias manifests in consistently inflated cosine similarity scores for same-platform pairs, with all differences being statistically significant ($p < 10^{-15}$) and effect sizes ranging from $d = 0.22$ to $d = 0.54$ (Table 4). These platform-specific variations indicate that effective bias mitigation requires tailored intervention strategies accounting for formal versus informal argumentation contexts.

6.3 Effectiveness of Targeted Fine-tuning

Finding 3: Fine-tuning on ArguBias improves overall embedding performance and independent benchmarks, while showing mixed effectiveness across specific bias types.

Fine-tuning demonstrates improvements across all models, with ArguBias test performance gains of 6.8-11.6 percentage points (Jina: +6.8pp, Nomic: +8.3pp, MPNet-v2: +11.6pp). Benefits extend across all platform combinations, with strong gains in same-platform scenarios: Congress-to-Congress (MPNet-v2: +20.7pp, Nomic: +12.7pp, Jina: +10.4pp) and Reddit-to-Reddit (MPNet-v2: +21.1pp, Jina: +11.7pp, Nomic: +11.6pp).

The benefits of fine-tuning generalize beyond training data topics, with average performance improvements of 7.1% on BWS and 5.4% on AFS benchmarks. These benchmarks contain arguments on topics absent from training data, suggesting fine-tuning enhances domain-general argument similarity performance.

Bias-specific analysis reveals mixed results. MPNet-v2 shows improvement on availability heuristic (-1.9pp) and affect heuristic (-2.3pp). However, lexical overlap bias degrades for MPNet-v2 (+3.1pp) and confirmation bias shows minimal improvements across models. This resistance reinforces our earlier finding that surface-level vocabulary matching represents the most fundamental vulnerability in embedding models. While fine-tuning can address complex structural biases, the systematic overconfidence introduced by shared terminology remains a challenge requiring alternative strategies (Table 5).

7 Conclusion

We introduce and evaluate the role that cognitive bias structures play in computational argumentation. We develop a novel dataset, ArguBias, consisting of 8,000 argument pairs from diverse sources and a taxonomy for cognitive bias structures that impair argument similarity judgments. Using this framework, we evaluate text embedding models on their vulnerability to cognitive biases, finding that lexical overlap bias causes severe similarity overestimation, while all models exhibit source-specific representational bias. Through minimal fine-tuning, embedding models achieve 6.8-11.6pp improvements on bias resistance with generalization to independent argument similarity benchmarks. This work provides the first systematic evaluation of cognitive biases in computational argumentation and establishes interventions for developing more transparent and robust argumentation systems.

Limitations

Dataset Scope. The ArguBias corpus is comprised of four controversial policy topics (abortion, gun control, nuclear energy, GMOs) in US political discourse. This limits generalizability to other domains, cultures and languages. Our analysis covers only English language arguments from two domains (Congressional hearings and Reddit), limiting representativeness across argumentative contexts. Also, some cognitive bias types have limited sample sizes (False equivalence: 9 pairs, Halo effect: 9 pairs), limiting statistical analysis of their effects. However, controversial topics ensures contexts where cognitive biases manifest systematically. Our cross-platform design also captures formal and informal discourse patterns prevalent in real-world argumentation systems.

Annotation Methodology. Our annotation process relies on ChatGPT-o3 for full-scale labeling with human validation of 100 pairs, achieving moderate inter-annotator agreement (Cohen's $\kappa = 0.471$). This agreement level reflects the inherent epistemological challenge of our task, cognitive bias structures fundamentally complicate positional similarity assessment. Disagreements cluster around the multidimensional topics of nuclear energy and GMOs, that resist binary classification, unlike traditionally polarized issues such as gun control or abortion.

Our deliberate focus on challenging boundary cases naturally produces moderate agreement consistent with other complex tasks, argument mining tasks achieving $k = 0.35$ - 0.63 (Lawrence and Reed, 2020), discourse relation annotation yielding $k < 0.65$ (Hoek et al., 2021), and opinion relation tasks reporting comparable moderate agreement (Brambilla et al., 2022). In preliminary evaluations on a subset of samples o3 achieved better human agreement ($k = 0.37$) over 4o ($k = 0.33$), leading us to select o3 for full-scale annotation. While this hybrid approach may introduce systematic biases, our validation confirms it captures the genuine complexity of cognitive bias effects where even expert human judgment encounters systematic challenges.

Methodological Constraints. Our framework employs cognitive bias structures as a post-hoc interpretability tool, providing observational insights into model behavior rather than causal explanations. The surface versus structural bias categorization, while theoretically motivated, is one possible taxonomy that could benefit from further refinement. However, this post-hoc approach enables immediate practical application by identifying vulnerabilities and actionable mitigation strategies.

Fine-tuning Scope. Our mitigation experiments examined three open-source embedding models with limited training (2 epochs), and evaluation focused on argument similarity benchmarks. The effectiveness of our approach across broader model architectures, training regimes, and downstream applications remains to be established. Nonetheless, consistent improvements across the three architectures and successful generalization to independent benchmarks demonstrate proof-of-concept that can guide broader implementation.

Argument Similarity Scope. Our work focuses on binary positional similarity judgments, which may not capture the full complexity of argumentative relationships. Real-world argumentation often involves degrees of alignment that our framework does not address. However, binary positional judgments directly serve critical applications like stance detection and misinformation identification, where nuanced similarity gradients often obscure essential positional distinctions.

Ethical Considerations

This work is in full compliance with the code of research ethics as established by ACM (Association for Computing Machinery, 2018) and adopted by ACL. We use only legally obtained, publicly available data without personally identifiable information.

Potential Risks. This work identifies systematic vulnerabilities in embedding models' that could enable adversarial manipulation of argument similarity systems. While our intention is to improve model robustness, as with any tool or knowledge, malicious actors could exploit these findings to undermine discourse integrity or automated declension-making processes.

Fairness Implications. Although our framework focuses on cognitive bias structures rather than social biases, these interactions could benefit marginalized communities by reducing the misclassifications and underrepresentation of their perspectives. We recognize that there are cultural norms and worldviews that are not equitably represented in argumentation systems, and careful consideration must be employed for deployment to avoid perpetuating inequalities or inadvertently privileging reasoning styles over others.

References

Association for Computing Machinery. 2018. Acm code of ethics and professional conduct. https://www.acm.org/code-of-ethics.

Libby Barak, Noe Kong-Johnson, and Adele Goldberg. 2019. Context effects on human judgments of similarity. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 135–137, Florence, Italy. Association for Computational Linguistics.

James H Barnes Jr. 1984. Cognitive biases and their impact on strategic planning. *Strategic Management Journal*, 5(2):129–137.

Jonathan Baron. 2023. *Thinking and deciding*. Cambridge University Press.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vincent Berthet. 2022. The impact of cognitive biases on professionals' decision-making: A review of four occupational areas. *Frontiers in psychology*, 12:802439.

Gianfranco Brambilla, Antonella Rosi, Francesco Antici, Andrea Galassi, Daniele Giansanti, Fabio Magurano, Federico Ruggeri, Paolo Torroni, Evaristo Cisbani, and Marco Lippi. 2022. Argument mining as rapid screening tool of covid-19 literature quality: Preliminary evidence. *Frontiers in public health*, 10:945181.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Tanise Ceron, Nico Blokker, and Sebastian Padó. 2022. Optimizing text representations to capture (dis) similarity between political parties. *arXiv preprint arXiv:2210.11989*.

André de Sousa and Karin Becker. 2023. Sssd: Leveraging pre-trained models and semantic search for semi-supervised stance detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 264–273.

Vahid Ghafouri, Jose Such, and Guillermo Suarez-Tangil. 2024. I love pineapple on pizza != I hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21046–21058, Miami, Florida, USA. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hans W. A. Hanley and Zakir Durumeric. 2024. Tata: Stance detection via topic-agnostic and topic-aware embeddings. *Preprint*, arXiv:2310.14450.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Tübingen, Germany. Association for Computational Linguistics.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2025a. A discourse analysis framework for legislative and social media debates. *17th ACM Web Science Conference*.

Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2025b. Wiba: What is being argued? a comprehensive approach to argument mining. In *Social Networks Analysis and Mining*, pages 337–354, Cham. Springer Nature Switzerland.

Brihi Joshi, Neil Shah, Francesco Barbieri, and Leonardo Neves. 2020. The devil is in the details: Evaluating limitations of transformer-based methods for granular tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3652–3659, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Edward J Joyce and Gary C Biddle. 1981. Anchoring and adjustment in probabilistic inference in auditing. *Journal of Accounting Research*, pages 120–145.

Dorian Jullien. 2016. All frames created equal are not identical: on the structure of kahneman and tversky's framing effects. *Œconomia. History, Methodology, Philosophy*, (6-2):265–291.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-based evaluation metrics by disentangling along linguistic factors. *arXiv preprint arXiv:2110.04399*.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. *Preprint*, arXiv:2309.17012.

Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. 2021. On classifying whether two texts are on the same side of an argument. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 10130–10138.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2024. Matryoshka representation learning. *Preprint*, arXiv:2205.13147.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2024. A comprehensive evaluation of cognitive biases in llms. *Preprint*, arXiv:2410.15413.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Amita Misra, Brian Ecker, and Marilyn A Walker. 2017. Measuring the similarity of sentential arguments in dialog. *arXiv preprint arXiv:1709.01887*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Nick Oh. 2024. In defence of post-hoc explainability. *arXiv preprint arXiv:2412.17883*.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Moritz Plenz, Philipp Heinisch, Anette Frank, and Philipp Cimiano. 2024. Pakt: Perspectivized argumentation knowledge graph and tool for deliberation analysis (with supplementary materials). *Preprint*, arXiv:2404.10570.

Sara Rajaee, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Looking at the overlooked: An analysis on the word-overlap bias in natural language inference. *arXiv preprint arXiv:2211.03862*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20.

Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng. 2024. Stance detection with explanations. *Computational Linguistics*, 50(1):193–235.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Paul Slovic, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. 2007. The affect heuristic. *European journal of operational research*, 177(3):1333–1352.

Keith E Stanovich, Maggie E Toplak, and Richard F West. 2008. The development of rational thought: A taxonomy of heuristics and biases. In *Advances in child development and behavior*, volume 36, pages 251–285. Elsevier.

Manveer Singh Tamber, Jasper Xian, and Jimmy Lin. 2024. Can't hide behind the api: Stealing black-box commercial embedding models. *Preprint*, arXiv:2406.09355.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Amos Tversky. 2003. *Preference, belief, and similarity: selected writings*. MIT Press.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*.

Robert B Zajonc. 1968. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Anna Zhdanko. 2019. Identification of cognitive manipulations that have the greatest impact on students in the internet. *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE)*, 7(1):35–42.

A Appendix

A.1 ArguBias Corpus Examples

Bias Type: False Equivalence
Argument 1: "If weapons of war have no place on our streets, neither do cars built for racing. No one making this argument also stops to think would I want to regulate cars in the way I am demanding that guns be regulated. Take the idea of a restriction on high capacity magazines, for example."
Argument 2: "A car does not even have to be street legal do you drive it at the race track or private land. Also there is no background check to buy a car and we do not stop felons and fugitives from justice, etc from buying cars. So I fail to see how they are more regulated then guns."

Figure 1: Both arguments appear to draw an analogy between cars and guns, which might lead someone to focus on this structural similarity and assume a shared conclusion or perspective. However, they actually arrive at different positions regarding regulation, making the similarity more superficial than substantive.

Bias Type: Affect Heuristic
Argument 1: "Numerous media accounts had analyzed in detail the possibility of creating dirty bombs from a combination of readily available commercial sources of radiation and common explosives. We all know that the credibility of these threats is inherently difficult to pinpoint. But we all also know of the seriousness with which these must be viewed and it has changed forever since September 11."
Argument 2: "Except for public overreaction based on the public irrational fear of radiation, dirty bombs are barely more dangerous than conventional bombs. More people would die from the explosion than would die from the radiation. Hell, more people might die from the toxic metal poisoning than would die from the radiation."

Figure 2: Argument 1 evokes the fear and seriousness associated with post-9/11 security threats, contributing to the perception of high risk. Argument 2, however, suggests that public fear is irrational and overrated. These shared references to public fear and risk might lead someone to mistakenly perceive the arguments as sharing similar concerns about dirty bombs, despite one emphasizing seriousness and the other downplaying the threat.

Bias Type: Anchoring Bias
Argument 1: "One would likely sink a submarine which are already nuclear powered and can stay out to sea as long as ship supplies hold out. They can even go underwater for long stretches at a time. So I see no advantage to putting a fusion reactor on a submarine."
Argument 2: "I learned somewhere that nuclear reactors on ships and submarines would not be as much of a problem as you would think, as the water surrounding a sunken sub or vessel is an excellent radiation shield"

Figure 3: Both arguments mention nuclear technology in the context of submarines, which can serve as an initial anchor. This shared topic might lead someone to perceive these arguments as more similar than they actually are, despite focusing on entirely different concerns, efficiency of fusion versus radiation safety.

Bias Type: Availability Heuristic
Argument 1: "A hundred Americans dying every day. So that means as we sit here today, there are Americans being killed by guns. I believe your statistics of 3 million people effectively being stopped through gun checks, and the universal background check bill is a bill of common sense."
Argument 2: "Probably annoying for collectors but I doubt the general public will be affected.Those are the only two areas that I think may be problems for some people. I really do not think the other things he is trying to do such as universal background checks are a bad idea. If anything it will end the demonization of firearm owners since people know they passed a background check."

Figure 4: Argument 1 uses a vivid image of 'a hundred Americans dying every day,' which might create strong emotional availability but lacks in Argument 2, which uses a more detached tone discussing the policy itself. This difference in emotional vividness could lead someone to perceive these arguments as more different than they actually are, despite sharing a core position supporting universal background checks.

Bias Type: Confirmation Heuristic
Argument 1: "Adoption wasn't the right decision for us because if I continued the pregnancy I would have wanted to parent, and the WIC, SNAP, and Medicaid programs aren't enough as it is."
Argument 2: "Adoption is always on the table if you do not want the kid yourself. Everyone wants an abortion at the time, but may regret it down the road. Down the road is when the mental health problems kick in."

Figure 5: Both arguments mention adoption, which might lead someone to focus on this shared element and perceive the arguments as more similar than they really are. Confirmation bias could cause someone to overlook their fundamentally different positions on the relationship between adoption and abortion.

Bias Type: Framing Bias
Argument 1: "For many years, there has been an assertion that abortion is safer than childbirth, and this has been used to defend the right to abortion."
Argument 2: "If you restrict abortion before addressing poverty and mental health conditions, then the second problem becomes worse and even harder to solve."

Figure 6: Both arguments are framed around the consequences of abortion or restricting abortion, although through different lenses (health vs. socioeconomic and mental health impacts). This might lead to an overestimation of their similarity because they both discuss consequences related to abortion, but they fundamentally focus on different issues.

Bias Type: Implicit Association Bias
Argument 1: "Some of them do not even give a shit, and that the worst part. A lot of them have no idea what they are voting for when they vote for these measures straight down partisan lines, they buy into the common sense refo room bullshit, because they do not understand anything about firearms. If they did, the assault weapons ban would not have been built around cosmetic features."
Argument 2: "For example, we are still pissed about the assault weapons ban of 9404, so when Feinstein tried that shit again in '13, we acted first and stopped it. Maybe they are are not that many antigun-ers as extreme as Feinstein. But if the antigun lobby ever wants to accomplish anything big, they need to reign in their radicals."

Figure 7: Both arguments use charged language associated with gun rights advocacy ('bullshit,' 'antigun lobby,' 'radicals'), which might activate implicit associations that lead someone to perceive these arguments as more similar in terms of general antigun sentiment, despite focusing on different entities within the antigun debate (voters vs. politicians).

Bias Type: Lexical Overlap Bias
Argument 1: "With that being said, yes Thorium is energy dense, yes we would reduce proliferation concerns using Thorium, yes it is technically a viable option for some countries. With that being said, Thorium is not an end all for nuclear power and there are a lot of issues with producing a steady flow of U233 from Th232 due to the long halflife of Pa233. Not only that, U232 a product of U233 is also highly radioactive making it very difficult to handle the fuel."
Argument 2: "Thorium is not fissile but it is fertile and can be converted to U233 in a reactor, same as U238 can be turned into Pu239. This fact means we have a virtually endless supply of nuclear fuel."

Figure 8: Both arguments use similar technical terminology related to thorium and nuclear reactions (e.g., 'U233,' 'fertile,' 'fissile'), which may lead someone to focus on these shared terms and overlook the fundamentally different conclusions and positions regarding the viability of thorium as a nuclear fuel.

Bias Type: Linguistic Distance Bias
Argument 1: "So in my view I am not sure what all this worry about storage is about. I think the correct way to deal with longlived nuclear waste is to burn it up in a fast reactor. Designs for moltensalt fast reactors are being developed by Moltex, Terrapower and others."
Argument 2: "I know it a bit late, but fast reactor neutrons from fission do not have anywhere near the neutron energy as neutrons from fusion. The median neutron energy from nuclear fission in fast reactors is .75 MeV, but in fusion the energy is closer to 14 MeV which can shatter anything heavier than bismuth. You can burn thorium directly without transmutation in this neutron spectrum, and you can burn all transuranic actinides as well.Mind this is all academic, since we do not have any controlled fusion reactors."

Figure 9: Argument 1 uses more general terms and a practical focus on waste management solutions, while Argument 2 delves into technical specifics about neutron energy in fast reactors. This difference in technical detail and complexity might lead someone to perceive these arguments as more different than they actually are, despite both supporting the concept of using fast reactor technology for nuclear waste management.